DOCUMENT RESUME

ED 181 028 TM 009 864

AUTHOR Koffler, Stephen I.
TITLE A Comparison of Approaches for Setting Proficiency Standards.
PUB DATE [79]
NOTE 34p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; *Comparative Statistics; *Cutting Scores; Decision Making; Elementary Secondary Education; *Mastery Tests; *Minimum Competency Testing; Multiple Choice Tests; State Programs; Statistical Analysis; Technical Reports; Testing Programs; Test Interpretation
IDENTIFIERS *Contrasting Groups Method; *Nedelsky Method; New Jersey

ABSTRACT

This research compared the cut-off scores estimated from an empirical procedure (Contrasting group method) to those determined from a more theoretical process (Nedelsky method). A methodological and statistical framework was also provided for analysis of the data to obtain the most appropriate standard using the empirical procedure. Data were provided from New Jersey's Minimum Basic Skills tests in reading and mathematics, administered in grades 3, 6, 9, and 11. (Tables show the degree of consistency between the proficiency standards estimated using the two procedures. Further, a modification of the linear discriminant function and the quadratic discriminant function, developed by Conover and Iman, is proposed as an appropriate method of analysis for the contrasting groups procedure). (Author/GDC)

A COMPARISON OF APPROACHES

FOR SETTING PROFICIENCY STANDARDS

STEPHEN L. KOFFLER

NEW JERSEY DEPARTMENT OF EDUCATION

A COMPARISON OF APPROACHES FOR SETTING PROFICIENCY STANDARDS

ABSTRACT

The purpose of this research is twofold. First, it compares the cut-off scores estimated from an empirical procedure (Contrasting Groups method) to that determined from a more theoretical process (Nedelsky method). Second, it provides a methodological and statistical framework for analysis of the data to obtain the most appropriate standard using the empirical procedure.

The results indicate that there was a consistency between the proficiency standards estimated using the two procedures. Further, a modification of the Linear Discriminant Function and the Quadratic Discriminant Function, developed by Conover & Iman (1978), is proposed as an appropriate method of analysis for the Contrasting Groups procedure.

A COMPARISON OF APPROACHES FOR SETTING PROFICIENCY STANDARDS

INTRODUCTION

The participants at four regional Minimum Competency Testing conferences sponsored by the Education Commission of the States defined minimum competency testing as "...tests (that) are constructed to measure the acquisition of competence or skills to or beyond a certain defined standard." (Miller, 1978).

The setting of such standards for these mandated minimum competency testing programs is one of the more important and controversial issues confronting the educational community (Linn, 1978). As Burton(1978) has noted, "...the setting of performance standards was suggested as a way to make educational decisions. The advocates of performance standards or criterion referenced testing in education wanted to base such decisions as high school graduation or the distributions of compensatory funds directly on performance measures."

As of July, 1978, thirty-seven states had implemented minimum competency testing programs, and in all remaining states, either legislative or administrative action is pending. This national movement toward minimum competency testing has brought about a new challenge to psychometricians and statisticians--the development of appropriate analytic procedures for the determination of proficiency standards. With the significance attached to the actual standard, it is imperative that an appropriate standard be determined, taking into consideration the difficulty of the assessment instrument, the skills to be addressed, and the expectations for success.

There are numerous methods to set standards.[1] Each method is essentially unique in its derivation; yet all possess a common quality--some degree of arbitrariness. The result of each procedure is based on judgment and the resulting standards are to some degree arbitrary. "All standard setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which performances are meritorious or acceptable and which are unacceptable or inadequate..." (Jaeger, 1976). However, because there does exist some degree of arbitrariness or judgment in all of the different standard setting procedures, it may be likely that the proficiency standard established might vary with the type of procedure used.

Andrew and Hecht (1976) found a substantial difference in the proficiency standard derived from two different methods. "It is perhaps not surprising that two procedures which involve different approaches to the evaluation of test items would result in different examination standards. Such examination standards will always be subjective to some extent and will involve different philosophical assumptions and varying conceptualizations."

Other than the Andrew and Hecht study, little empirical research has been reported that compares proficiency standards obtained using different procedures. However, this area is indeed a critical one, especially when considering the many states and local school districts mandating minimum competency testing programs, and with them, an associated, defensible proficiency standard. It is important to determine a standard that is both reliable and consistent, regardless of the procedure used to determine it.

---

[1] See, for example, Hambleton & Eignor, 1978a, 1978b; Hambleton, et. al., 1978; Jaeger, 1976; Millman, 1973; Meskauskas, 1976; Shepard, 1976.

One of the most important results of the Andrew and Hecht study was the high level of agreement in the determination of the proficiency standard using the same method across two teams of judges (Hambleton, 1978). It appears that, based on that study, one can obtain reliable and consistent standards using the same technique at different points in time or with different judges. However, the comparability among methods has not been fully investigated.

The research described in this paper is an attempt to determine the comparability of proficiency standards developed using different procedures. It compares the proficiency standard determined from an empirical procedure based upon judgments about the mastery status of a sample of test takers (Contrasting Groups method) to the proficiency standard determined from the more usual procedure based upon judgments about the questions on the test (Nedelsky method).

The paper further illustrates the methodology underlying both types of procedures and proposes a nonparametric statistical procedure to determine the proficiency standard using the Contrasting Groups procedure.

BACKGROUND

In September, 1976, New Jersey Governor Brendan T. Byrne signed into law a measure establishing uniform statewide minimum proficiency requirements in computation and communications skills for all publicly educated elementary and secondary students. To implement this law, the New Jersey State Board of Education directed that Minimum Basic Skills (MBS) tests in reading and mathematics be developed and implemented commencing with the 1977-1978 school year. The State Board of Education required that a mastery level be set for each test and that the tests be administered to all students (excluding

students classified as handicapped or Limited English Speaking ability) in grades three, six, nine, and eleven in the spring of each school year.

One of the major purposes of the MBS Tests is the identification of students who may require remediation in the basic skills. Because the MBS Tests are a screening device, it was decided that it would be most appropriate to set a cut-off score based on a total score for each test.[2]

Zieky and Livingston (1977) suggested four methods for the standard setting process, two requiring judgments about the questions on the test (the Nedelsky (1954) and Angoff(1970) methods) and two requiring judgments about the mastery status of a sample of students taking the test (the Borderline Group and Contrasting Group methods). A modification to the Nedelsky approach was ultimately decided upon for the determination of the statewide proficiency as being the most efficacious for New Jersey's particular situation. (Meissner, 1977).

NEDELSKY PROCEDURE

Zieky and Livingston (1977) describe the Nedelsky standard setting procedure as one "...based on judgments of what certain types of students should be able to do...Nedelsky's method estimates the expected average score of a group of students with minimally acceptable performance. Students whose performance is not acceptable are likely to score below that level, and students whose performance is acceptable are likely to score above that level..."

Nedelsky's procedure can only be used with multiple choice tests, since it requires a judgment about each possible distractor to each item.[3] The use of the procedure requires that a panel of experts be assembled to

[2] The cut-off score is that point on the test score scale that is used to sort examinees into two categories that reflect different levels of proficiency relative to a particular objective measured by the test. (Hambleton, 1978).

[3] All of the New Jersey Minimum Basic Skills tests are composed of four choice items.

examine each item.[4] For each of the items, every panelist eliminates those distractors that he/she is confident that a student with a minimal mastery will not consider as possible correct answers to the item. The reciprocal of the remaining choices is the estimated probability that the minimal master will correctly answer the item. The sum of the estimated item probabilities is the score that the panelist expects the minimal master to attain. This score is the panelist's best estimate of the proficiency standard. The overall best estimate for the proficiency standard is the mean of the panelists' estimates.

The following instructions were given to the panelists charged with developing the New Jersey minimum proficiency standards:[5]

1.) The first step in applying the standard setting procedure is to think about what you consider to be the lowest level of performance you are still willing to classify as mastery of the skills measured by the test that you worked on. If you have recent classroom experience, it may help you to think about students you have known that were just barely good enough to be considered masters of the basic skills measured by the test.

We expect that there will be some differences of opinion as to what is meant by minimally acceptable performance.

2.) The second step is to look at the first question in the test and decide how many wrong answers are so wrong that even the minimally acceptable student would know that they are wrong.

For example, the following question is similar to one on the Grade Three Math test:

> The school lunchroom served 506 people on Monday and 315 people on Tuesday. How many people were served on the two days?
>
> (A) 191
> (B) 201
> (C) 811
> (D) 821

---

[4]The panelists were the members of the committees established for the purpose of test development. Most panelists were experts in reading and mathematics from local New Jersey school districts. The remainder were from higher education or business. There were four separate committees, each with different panelists.

[5]Communications from Dr. Michael Zieky to the panelists.

You may decide that even the minimally competent student should know that A and B are wrong because the total for two days would be greater than the number on any single day. But, you may decide that wrong answer C involves an error that the minimally competent student would not know is wrong. You would therefore decide that two wrong answers for the questions are so wrong that even the minimally competent students would know that they are wrong.

3.) We will then ask for a few volunteers to tell the group which wrong answers were selected and their reasons for selecting them. You will be encouraged to discuss the choices. The discussions may either confirm your earlier opinions or change your mind.

4.) The last step is for you to record the number of wrong answers you selected as being so wrong that even the minimally qualified student would know they are wrong.

5.) We will go on to the next question and repeat the process. After you are done, we will estimate the tentative standard for each test based on the data you provided.

RESULTS OF THE NEDELSKY PROCEDURE

Table 1 lists the mean estimated proficiency standards for each test both as a raw score and as a percentage of the test items that resulted from using the Nedelsky procedure. Also presented in Table 1 are the number of panelists who participated in the process, the number of items on the tests, the minimum and maximum estimated standards and the standard deviation of the estimates. The minimum and maximum estimated standards and the standard deviation of the estimates are shown to indicate the level of agreement among the panelists for a given test.

With the exeption of the estimates for the ninth and eleventh grade mathematics tests, the mean estimated cut-off scores for the other tests were somewhat consistent, ranging from 52.5%-81.9% correct for reading and 58.2%-65.9% correct for mathematics. The estimates for the ninth and eleventh grade mathematics tests were 37.2% and 37.3% correct, respectively.

TABLE 1

ESTIMATED CUT-OFF SCORES USING THE NEDELSKY PROCEDURE

TEST	N ITEMS	AVERAGE RAW SCORE STANDARD	AVERAGE PERCENTAGE STANDARD	N COMMITTEE MEMBERS	MINIMUM RAW SCORE ESTIMATE	MAXIMUM RAW SCORE ESTIMATE	STANDARD DEVIATION OF RAW SCORE ESTIMATE
READING 3	100	63.6	63.6%	10	53.8	74.2	8.05
READING 6	95	49.9	52.5	10	36.9	58.1	7.17
READING 9	110	79.8	72.5	9	70.1	85.6	6.86
READING 11	110	90.1	81.9	8	80.7	95.7	5.66
MATHEMATICS 3	100	58.2	58.2	10	46.7	66.6	6.55
MATHEMATICS 6	100	65.9	65.9	10	56.9	74.1	6.09
MATHEMATICS 9	95	35.3	37.2	7	30.0	38.8	3.05
MATHEMATICS 11	90	33.6	37.3	7	28.7	36.3	2.61

The panelists responsible for those tests believed that their estimates were an appropriate representation of the Nedelsky procedure. However, they were also of the opinion that these estimates were not realistic for use as a statewide minimum proficiency standard.[6]

CONTRASTING GROUPS METHOD

Although the actual statewide minimum proficienoy standards were established using the Nedelsky procedure, it was decided that a valuable contribution could be made by utilizing another standard setting method in a research situation and comparing the results of the two procedures. The most appropriate procedure for these purposes was determined to be the Contrasting Groups method.

Whereas the Nedelsky procedure requires judgments about test items in relation to the theoretical "minimal master", the Contrasting Groups approach requires judgments about the mastery levels of actual test takers and is based upon empirical evidence. Zieky and Livingston(1977) state that "...testing specialists who favor these (Contrasting Groups) methods believe that judgments about students are usually more meaningful than judgments about test items. People in our society (and teachers in particular) are accustomed to judging the performance or achievement of other people as being adequate or inadequate for some purpose. The method discussed earlier... (i.e., the Nedelsky...technique) ... present(s) the judges with a rather unfamiliar task, such as imagining a "minimally acceptable" student and guessing the probability that the student will answer a given question correctly. In contrast, the ...(Contrasting Groups method)...require(s) a much more common

[6] That committee imposed another method to estimate the standards. (Meissner, 1977). For the purposes of imposing the statewide proficiency standards, the precise mean estimates derived from the Nedelsky procedure were not used. Rather, it was decided that the standards would be 65% correct in mathematics and 75% in reading. However, for the purposes of this study, the actual Nedelsky results will be used.

type of judgment: the judgment of a real student's level of mastery as adequate (or) inadequate..."

The Contrasting Groups method sets a standard at that test score that best separates those students judged to be masters from those students judged to be nonmasters (Zieky and Livingston, 1977). To utilize the Contrasting Groups method, one must determine the mastery/non-mastery status and the test score for a representative sample of the test takers.

The following procedure in relation to the Contrasting Groups procedure was used during the administration of the New Jersey Minimum Basic Skills (MBS) tests in April, 1978. The students' teachers were asked to voluntarily indicate on each student's answer sheet whether the teacher judged that the student did have a minimal mastery of the skills addressed on the tests. The instructions given to the teachers were as follows:

> To assist the Bureau of Research and Assessment in obtaining information relative to the state proficiency standards, examiners are requested to provide the following information on each student's answer sheet. In the box under the words "For Teacher Use Only" on Side 1 of each student's answer sheet there are letters "R" and "M". Next to each of these letters are two ovals. One of the ovals has a "Y" for "yes" and the other oval has an "N" for "no". Examiners are asked to grid the "Y" ovals if, in the judgment of that student's subject matter teachers, he/she is minimally proficient in reading and mathematics. If the student is not minimally proficient in reading and mathematics, the "N" ovals should be gridded. (New Jersey Educational Assessment Program Minimum Basic Skills Tests District Test Coordinator manual, 1978).

RESULTS OF THE CONTRASTING GROUPS METHOD

Table 2 indicates the number of students for which mastery/non-mastery judgments were obtained and the total number of students tested in April, 1978. From Table 2, it is readily apparent that many more judgments were rendered about elementary students than secondary students.

TABLE 2

MASTERY/NON-MASTERY DATA FOR THE APRIL 1978 ADMINISTRATION OF THE MBS TESTS

TEST	TOTAL STUDENTS TESTED	TOTAL STUDENTS WITH MASTERY/ NON-MASTERY DATA	PERCENT OF STUDENTS WITH MASTERY/NON-MASTERY
READING 3	90229	44255	49.0%
READING 6	95848	45395	47.4
READING 9	109446	19573	17.9
READING 11	98214	16447	16.7
MATHEMATICS 3	90183	44131	48.9
MATHEMATICS 6	95736	45381	47.4
MATHEMATICS 9	108531	19601	18.1
MATHEMATICS 11	97631	15761	16.1

Before analyzing the data, it was necessary to determine if the sample of students for which mastery information was available was representative of the population of test takers in New Jersey. All of the 611 school districts in New Jersey have been assigned to one of twelve categories called District Factor Groups (DFG), based on the relative socio-economic status of the residents of the school districts and to one of four groups based on geographic location. Categorizing each district into one of the forty eight strata, determining the percentage of the population in each grade and the percentage of students from the sample in each stratum, it was determined that the sample was not representative.

It is, however, crucial that the sample be representative of the intended population to determine an appropriate proficiency standard using the Contrasting Groups method. A stratified random sampling plan was utilized, considering the relative percentage of students in each of the strata, to obtain a representative sample of the population. Table 3 lists the resulting number of students for each test who comprised the sample following the sampling procedure. For the third and sixth grades, the sample represented approximately 33% of the population, while for the ninth and eleventh grades, the sample was approximately 10% of the population. Table 3 also lists the number (and percent) of masters and non-masters in the sample.

Statistically, the determination of the appropriate cut-off score, given the mastery/non-mastery information and the students' test scores, is a well defined process. The model is an example of the two group univariate classification problem. There are two groups (masters and non-masters) and a student must be rationally assigned to one of these two groups, based on

TABLE 3

MASTERY/NON-MASTERY DATA FOR THE STRATIFIED RANDOM SAMPLE OF STUDENTS

TEST	TOTAL STUDENTS WITH MASTERY/NON-MASTERY DATA	NUMBER OF MASTERS	PERCENT OF MASTERS	NUMBER OF NON-MASTERS	PERCENT OF NON-MASTERS
READING 3	31126	24373	78.3%	6753	21.7%
READING 6	30912	23494	76.0	7418	24.0
READING 9	11390	8877	77.9	2513	22.1
READING 11	9910	8406	84.8	1504	15.2
MATHEMATICS 3	31032	24336	78.4	6696	21.6
MATHEMATICS 6	30921	23035	74.5	7886	25.5
MATHEMATICS 9	11417	8903	78.0	2514	22.0
MATHEMATICS 11	9488	7933	83.6	1555	16.4

his/her test score. Thus, that test score must be determined such that a student scoring below the score will be properly classified into the non-master group and a student scoring at or above the score will be properly classified into the master group.

Classically, Fisher's Linear Discriminant Function(LDF) (1936) has been used to obtain solutions to these types of problems (i.e. to determine group membership). Welch (1939) provided theoretical background for Fisher's LDF by adapting the hypothesis testing arguments of Neyman and Pearson. Welch showed that classification procedures were equivalent to obtaining appropriate mutually exclusive and exhaustive regions R_i of the sample space. For a specific partitioning of the sample space into such regions R_i, an observation to be classified (i.e., a student categorized either as a master or a non-master) is assigned to population S_i (i.e., the mastery or non-mastery population) if the p-dimensional data point (i.e., the one dimensional test score) lies in region R_i.

Using the rationale of Welch, there are two possible errors that can be committed:

1. A student can be classified as a master when the student has not adequately mastered the objectives (i.e. the experts judged him/her to be a non-master; however, the classification procedure classified him/her as a master).

2. A student can be classified as a non-master when the student has adequately mastered the objectives (i.e., the experts judged him/her to be a master; however, the classification procedure classified him/her as a non-master).

Zieky and Livingston (1977) refer to the first type of error as a "false master" and to the second type of error as a "false non-master".

Associated with each of these two types of errors is a probability of committing the error (called the probability of misclassification). Welch

showed that the optimum classification procedure is one which partitions the sample space in such a manner that the corresponding probabilities of misclassification are minimized.

The form of the optimal classification procedure is the ratio of the known density functions. (Welch, 1939). The observation (or student) to be classified (either as a master or a non-master) is classified into population S_1 (i.e., the master population) if the value of the likelihood ratio is greater than some appropriately determined constant k; the observation is assigned to population S_2 (i.e., the non-master population) if the value of the likelihood ratio is less than the constant k.[7] Furthermore, the constant k defining the decision rule for group membership (or mastery/non-mastery status) can be established based on the a priori probabilities of group membership, if known, and the relative "costs" of misclassification.

The constant k is defined as follows:

$$k = \frac{C_{12}q_2}{C_{21}q_1} \qquad (1)$$

where C_{ij} is the "cost" of misclassifying an observation actually belonging to S_j into S_i $(i,j = 1,2)$, and q_i $(i = 1,2)$ is the a priori probability of group membership.

For the Contrasting Group procedure, the a priori probabilities of group membership are estimated by the proportion of examinees in the sample who are judged to be masters (q_1) or non-masters (q_2). Further, unless the "costs" of each type of classification error can be deduced, it is customary to assume equal "costs" of misclassification. Hence, for the Contrasting

[7] When the value of the likelihood ratio equals k, it is customary to use a randomized procedure to classify the observation. However, for the purposes of minimum competency testing, it is necessary to uniformly classify a student whose score is equal to the cut-off either as a master or a non-master. For the purposes of this paper, a student whose score is equal to the cut-off (k) will be classified as a master.

Groups application,

$$k = \frac{q_2}{q_1} \qquad (2)$$

where q_i $(i = 1,2)$ is the proportion of masters or non-masters in the sample.

If it can be determined that one type of error is more serious than the other, the constant k can be adjusted and the standard set accordingly so as to minimize that type of error, at the cost of making more errors of the other type. For example, if it is more "costly" to obtain false masters, then C_{12} would be greater than C_{21} and, hence, k increases. Therefore, the resulting standard would be higher and effectively the number of false masters would decrease as a result. For purposes of this research, it was assumed that the relative "costs" of misclassification were equal. Hence, the constant k was determined as in (2).

When both populations are normally distributed with equal variance-covariance matrices and known mean vectors, the ratio of the density functions becomes Fisher's LDF. When the population parameters are unknown, Hoel and Peterson(1949) and Fix and Hodges(1951) have shown that the form of the optimal classification procedure is the ratio of the density functions with the sample statistics replacing the population parameters. Anderson (1951) developed a statistic (W) which is the sample based analogue to the LDF. Thus, by assuming that the population of masters' and non-masters' test scores are normally distributed with equal but unknown variances and unknown means, then the optimal classification procedure (or best method by which to determine the proficiency standard) is Anderson's W statistic. Given the above assumptions, a student would be classified as a master if

$$\{(\bar{X}_1 - \bar{X}_2)/S^2\}\{X - (\bar{X}_1 + \bar{X}_2)/2\} \geq \text{Ln}(q_2/q_1) \qquad (3)$$

and as a non-master if

$$\{(\bar{X}_1 - \bar{X}_2)/S^2\}\{X - (\bar{X}_1 + \bar{X}_2)/2\} < \text{Ln}(q_2/q_1) \qquad (4)$$

where $\bar{X}_1$ is the sample mean of the masters' test scores, $\bar{X}_2$ is the sample mean of the non-masters' test scores, S^2 is the pooled sample variance, X is the test score of the student to be classified, and q_i $(i = 1,2)$ is the a priori probabilities of group membership. Thus, using Anderson's W as above, the proficiency standard is the smallest test score such that equation (3) is true.

Gessaman and Gessaman(1972) have shown that the LDF is not robust to departures from the equality of the variance-covariance matrix assumption. When the populations are multivariate normal, but the variance-covariance matrices are not equal, the most appropriate statistic to use is the Quadratic Discriminant Function (QDF), which is the ratio of the density functions, assuming unequal variance-covariance matrices.

Using the QDF, a student would be classified as a master if

$$X\left(\frac{\bar{X}_1}{S_1^2} - \frac{\bar{X}_2}{S_2^2}\right) - \frac{X^2}{2}\left(\frac{1}{S_1^2} - \frac{1}{S_2^2}\right) - \tfrac{1}{2}\left(\frac{\bar{X}_1^2}{S_1^2} - \frac{\bar{X}_2^2}{S_2^2}\right) + \tfrac{1}{2}\text{Ln}(S_2^2/S_1^2) \geq \text{Ln}(q_2/q_1) \qquad (5)$$

and as a non-master if

$$X\left(\frac{\bar{X}_1}{S_1^2} - \frac{\bar{X}_2}{S_2^2}\right) - \frac{X^2}{2}\left(\frac{1}{S_1^2} - \frac{1}{S_2^2}\right) - \tfrac{1}{2}\left(\frac{\bar{X}_1^2}{S_1^2} - \frac{\bar{X}_2^2}{S_2^2}\right) + \tfrac{1}{2}\text{Ln}(S_2^2/S_1^2) < \text{Ln}(q_2/q_1) \qquad (6)$$

where $\bar{X}_1$ and S_1^2 are the sample mean and variance, respectively, of the masters' test scores, $\bar{X}_2$ and S_2^2 are the sample mean and variance, respectively, of the non-masters' test scores, X is the test score of the student to be

classified, and q_i (i = 1,2) is the a priori probabilities of group membership. Using the QDF, the proficiency standard is the smallest test score such that equation (5) is true.

If the test scores were distributed according to univariate normal distributions, then the problem of determining the optimal proficiency standard is solved by using either the LDF or QDF. However, it is unlikely that a set of minimum basic skills test scores is representative of a normal distribution. Due to the nature of these types of tests, the scores are usually skewed to the left. Kolmogorov-Smirnov tests showed that the data utilized in this study were not representative of normal distributions ($p < .01$).

Lachenbruch, Sneeringer and Revo (1973) & Koffler and Penfield (1979) have shown that the LDF and QDF are not robust procedures when used to classify observations from non-normal distributions. Thus, it is not appropriate to use the LDF or QDF when classifying observations that are drawn from non-normal populations. When the samples are drawn from known distributions, the form of the optimal procedure is the ratio of the density functions. In most instances, however, information is obtained from samples drawn from unknown populations and alternative procedures must be employed to estimate the density functions.

Several nonparametric procedures have been compared empirically by Gessaman and Gessaman(1972) & Koffler and Penfield(1979). Procedures such as the Nearest Neighbor with Probability Blocks and the Loftsgaarden-Quesenberry density estimator (1968) have been shown to effectively classify observations from non-normal distributions. However, these procedures are generally difficult to use and computer programs are not readily available.

Conover and Iman (1978) have suggested another approach based on the use of a mathematical transformation on the data so that each distribution function is approximately normal. According to Conover and Iman, once the transformation is applied to the data, then the LDF or QDF can be utilized. Lachenbruch (1975) & Moore and Smith(1975) have shown that a rank transformation is an appropriate one for all distributions. Conover and Iman (1978) empirically contrasted results obtained using the rank transformation procedure with those from other procedures. They concluded that if the data are normally distributed, the rank method performs essentially as well as the LDF and QDF methods; if the data are non-normal, the method works as well as any of the nonparametric methods investigated.

Based on the study by Conover and Iman, it is appropriate to use the LDF or QDF classification statistic to determine the proficiency standard, with ranks replacing the corresponding test scores. In general, the ranking procedure replaces each p-dimensional sample value by its corresponding rank, from rank 1 for the smallest value to rank N for the largest value for the combined samples ($N = n_1 + n_2$, where n_i, $i = 1,2$, is the sample size of the ith sample). Each of the p dimensions is ranked independently. Then, once the ranking is completed, the LDF or QDF is applied to the ranks.

For the purposes of the Contrasting Groups standard setting determination, the following procedure was used. For each of the eight sets of data, a ten percent completely random sample was drawn. Then, for each test, first the test scores for both groups of students were combined and replaced by their ranks, from rank 1 for the smallest test score to rank N for the largest score in the combined sample. When more than one student had the same test score, a midranks procedure was used to assign the ranks. Once the ranks were assigned, the appropriate sample means and variances were determined.

The variances were analyzed using the Mood test for scale to determine if a significant difference existed between the two samples. For all eight situations, a significant difference did exist ($p < .01$). Therefore, it was decided to use the ranked analogue to the QDF to determine the proficiency standard rather than the LDF.

The mean ranks and variances of the ranks were applied to equations (5) and (6) to determine that rank which best separated the master group from the non-master group, assuming equal "costs" of misclassification and a priori probabilities of group membership estimated from the proportion of masters and non-masters in the samples (see Table 3). Once the smallest rank was determined which satisfied equation (5), it was then transformed to the original test score. This test score represented the optimal proficiency standard, based on the Contrasting Groups method.

Table 4 summarizes the proficiency standards derived from the ranked QDF method and from the Nedelsky method. Table 4 also presents the percent of false masters and false non-masters that resulted from using the proficiency standards derived from each method, and the percent of students, in the population, who scored below these test scores.

RESULTS

The most striking result from Table 4 concerns the lack of a Contrasting Groups cut-off score estimate for the eleventh grade mathematics test. Given the relationship among the test scores, the mastery/non-mastery judgments and the a priori probabilities of group membership, every rank satisfied equation (5). Hence, using this procedure, all students would have been classified as masters, regardless of their test score or their teacher's judgment.

TABLE 4

A COMPARISON OF THE NEDELSKY AND CONTRASTING GROUPS RESULTS

TEST	NEDELSKY STANDARD	PERCENT FALSE MASTERS	PERCENT FALSE NON-MASTERS	CONTRASTING GROUPS STANDARD	PERCENT FALSE MASTERS	PERCENT FALSE NON-MASTERS	NEDELSKY PERCENT BELOW STANDARD[1]	CONTRASTING GROUPS BELOW STANDARD[1]
READING 3	64	76.9%	3.3%	72	65.5%	5.7%	7.4%	11.6%
READING 6	50	86.8	1.2	69	50.2	10.7	4.4	20.1
READING 9	80	49.4	11.8	80	49.4	11.8	20.2	20.2
READING 11	90	45.6	11.7	50	93.0	0.9	17.7	1.3
MATHEMATICS 3	58	55.9	8.8	55	61.1	7.1	17.1	14.3
MATHEMATICS 6	66	30.9	20.0	61	44.3	12.8	31.2	23.5
MATHEMATICS 9	35	96.0	0.6	58	51.1	11.1	1.7	19.4
MATHEMATICS 11	34	99.3	0.6	0	100.0	0.0	0.8	0.0

[1] The figure represents the percent of students, in the population of test takers, who scored below the respective cut-off score. See Table 2, p. 10, for the total number of students who were administered each test.

This result is clearly not acceptable for use with a minimum proficiency test. The result suggests three fundamental questions: 1) Is the discriminant analysis procedure inappropriate for determining proficiency standards? 2) Does the eleventh grade mathematics test discriminate between masters and non-masters (i.e., is it appropriate as a minimum competency instrument)? 3) Are the teachers' judgments valid in determining the mastery/non-mastery status of the eleventh grade students?

Considering the first question, it should be clear that the unusual result for the eleventh grade mathematics test was a function of the extreme overlap of the masters' and non-masters' distributions, not an inconsistency in the statistical procedure itself. Whenever there is overlap between distributions, there is the question of whether there are two distributions or only one. As a result, a great deal of misclassification is to be expected.

Anderson(1951) has shown that the Linear Discriminant Function is normally distributed with mean and variance, as well as probability of misclassification, a function of the Mahalanobis distance between the two populations. The greater the separation between the two distributions (i.e. the larger the Mahalanobis distance), the better the classification and the more easily discernible the cut-off score for the Contrasting Groups procedure. Conversely, as the Mahalanobis distance between the two distributions decreases, the more difficult the ability to correctly classify the observations, and the less appropriate the cut-off score.

Anderson(1951), further, has shown that, assuming that the two distributions are multivariate normal with known parameters, then the probability of misclassification is given by

$$\Phi(-\{c + \alpha/2\}/\{\alpha\}^{\frac{1}{2}}) \qquad (7)$$

$$\Phi(\{c - \alpha/2\}/\{\alpha\}^{\frac{1}{2}}) \qquad (8)$$

where c = Log (k), α is the Mahalanobis distance between the two distributions, and $\Phi(X)$ represents the value of the normal distribution function at point X. When the population parameters are unknown, the sample statistics can be used to provide an estimate of the expected probabilities of misclassification.

Table 5 lists the Mahalanobis distance between the two distributions, the value of the constant Log (k), and the estimated probability of misclassification for making a false master decision and a false non-master decision. Equation (7) denotes the probability of making a false master decision, while equation (8) is the probability of making a false non-master decision. Examining Table 5, it is evident that the largest estimated probability of obtaining a false master and the smallest estimated probability of obtaining a false non-master is associated with the eleventh grade mathematics test. These facts reinforce the notion of overlap of the distributions, given the a priori probabilities of group membership, and support the establishment of a very low cut-off score.

To address the issue of the validity of the eleventh grade mathematics test as a minimum competency testing instrument, one must consider the test development process. Content specialists were assembled to develop skill specifications and to review the resulting test items. Further, minority reviews were conducted to determine any bias in the instruments. Because of the work of the test development committees and the empirical results of the field test, it must be assumed that the eleventh grade mathematics test is a valid instrument and may be used to discriminate between masters and non-masters.

The question concerning the validity of the judgments about the students' mastery/non-mastery status is important to consider. There are many students who drop out of school between the ninth and eleventh grades. It is generally assumed that those who drop out are the poorest achieving students. A

TABLE 5

ESTIMATED PROBABILITIES OF FALSE MASTERS AND FALSE NON-MASTERS

TEST	MAHALANOBIS DISTANCE	LOG (k)	FALSE MASTER PROBABILITY	FALSE NON-MASTER PROBABILITY
READING 3	1.2653	-1.2832	0.7190	0.0446
READING 6	1.6319	-1.1525	0.6026	0.0618
READING 9	1.7113	-1.2599	0.6217	0.0526
READING 11	1.6071	-1.7190	0.7642	0.0239
MATHEMATICS 3	1.3425	-1.2891	0.7019	0.0455
MATHEMATICS 6	1.6887	-1.0720	0.5714	0.0708
MATHEMATICS 9	1.7168	-1.2657	0.6217	0.0526
MATHEMATICS 11	1.3634	-1.6287	0.7910	0.0244

question which must be raised concerns whether the teachers would have judged the same students as non-masters if the dropouts were also included in the samples. Further, the effect of the inclusion of the dropouts must be considered. It is hypothesized that the dropouts would have been judged to be non-masters and also would generally obtain test scores at the lower tail of the distribution. This would have the effect of causing a greater separation between the two distributions which would have resulted in a more appropriate cut-off score.

Comparing the other results from Table 4, it is evident that the standard obtained by the Contrasting Groups procedure was substantially larger than that obtained by the Nedelsky method for two tests (grade 6 reading and grade 9 mathematics); the Contrasting Groups procedure's cut-off score was slightly larger in one test (grade 3 reading); the Nedelsky estimate was slightly greater for two of the tests (grade 3 and 6 mathematics); the Nedelsky cut-off score was substantially larger in the eleventh grade reading test; and, there was complete agreement between the two estimates in the ninth grade reading test.

Despite this seeming inconsistency between the two procedures' results, a number of patterns do emerge. First, the large inconsistency in the standards for the eleventh grade reading test may be attributed to the dropout hypothesis raised in relation to the eleventh grade mathematics test. Further, the inconsistency between the ninth grade mathematics standards may be a function of the inappropriateness of the committee's Nedelsky approach. The results of the sixth grade reading test remain troublesome. There was a great discrepancy between the cut-off scores; however, there appears to be no plausible explanation to discount either of the estimates.

Excluding the ninth grade mathematics test, the eleventh grade reading test and the sixth grade reading test on the basis of explanations for the aberrent results, there was a consistency in the estimates for the other grades. The discrepancy between the estimates ranged from no difference (grade 9 reading) to an eight point raw score difference (grade 3 reading).

CONCLUSIONS

The major conclusion to be drawn from this study is that there is agreement between the cut-off scores developed by the Nedelsky and the Contrasting Groups procedures. Based on the results of this study, consistent proficiency standards can be developed, regardless of whether the theoretical Nedelsky approach or the more empirically based Contrasting Groups procedure is applied. Further, for the areas in which there was no agreement, there were plausible explanations for the discrepancies, with the exception of the sixth grade mathematics test.

The fact that there were tests for which the two cut-off scores were not consistent leads to a recommendation that there should not be a reliance upon one standard setting procedure to determine cut-off scores; rather, a number of procedures should be utilized. Further, a careful analysis of the data, judgments, and extraneous conditions which may be present in a particular situation and which may affect the estimates for a given procedure, must be considered in determining which cut-off score estimate will be selected as the ultimate proficiency standard. In the present study, if only the Nedelsky procedure was utilized, inappropriate standards would have been generated for both the ninth and eleventh grade mathematics tests; if only the Contrasting Groups procedure was utilized, inappropriate standards would have been selected for the eleventh grade reading and mathematics tests.

The final conclusion that should be drawn from the present study concerns the statistical analysis of the Contrasting Groups data. It is suggested that the most efficacious and statistically sound procedure to use for empirically developing the cut-off score is the univariate analogue to the Linear Discriminant Function or the Quadratic Discriminant Function in which the ranks of the scores replace the actual data.

Further research is suggested by these results. It would be valuable to determine if the consistencies discovered between the Nedelsky and Contrasting Groups procedures extend to other methods. A study should be attempted in which the Nedelsky, Contrasting Groups, Borderline Groups, and Angoff methods are compared. Further, in relation to the Contrasting Groups and Borderline Groups methods, research is needed to test the validity of the expert judgments and to determine more precise judgmental methods.

REFERENCES

Anderson, T.W., Classification by multivariate analysis. Psychometrika, 1951, 16, 31-50.

Andrew, B.J. & Hecht, J.T., A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W.J., Scales, norms, and equivalent scores. In R.L. Thorndike (ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971, 514-516.

Burton, N.W. Societal Standards. Journal of Educational Measurement. 1978, 15, 263-272.

Conover, W.J. & Iman, R.L., The rank transformation as a method of discrimination with some examples. Sandia Laboratories, 1978.

Fisher, R.A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7, 179-188.

Fix, E. & Hodges, J.L., Nonparametric discrimination: consistency properties. Lackland Air Force Base, Texas: U.S. School of Aviation Medicine, 1951.

Gessaman, M.P. & Gessaman, P.H., A comparison of some multivariate discrimination procedures. Journal of the American Statistical Association, 1972, 67, 468-472.

Hambleton, R.K., On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R.K. & Eignor, D., A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, Mass: School of Education, University of Massachusetts, 1978(a).

Hambleton, R.K & Eignor, D., Determining minimum competency levels: A consideration of issues, methods, and implementation strategies. A paper presented at the AERA Conference on Minimal Competency Testing, Washington, D.C., 1978(b).

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B., Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hoel, P.C. & Peterson, R.P., A solution to the problem of optimal classification. Annals of Mathematical Statistics, 1949, 20, 433-438.

Jaeger, R.M., Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 1976, 18, 22-27.

Koffler, S.L. & Penfield, D.A., Nonparametric discrimination procedures for non-normal distributions. Journal of Statistical Computation and Simulation, 1979, 8, 281-299.

Lachenbruch, P.A., A problem in discrimination using ranks. Paper presented at the 8th annual Symposium on Interface of Computer Science and Statistics, Los Angeles, 1975.

Lachenbruch, P.A., Sneeringer, C. & Revo, T., Robustness of the linear and quadratic discriminant function to certain types of non normality. Communications in Statistics, 1973, 1, 39-57.

Loftsgaarden, D.O. & Quesenberry, C.P., A nonparametric estimate of a multivariate density function. Annals of Mathematical Statistics, 1965, 36, 1049.

Linn, R.L.,Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 1978, 15, 301-308.

Meissner, F.G., State of New Jersey Minimum Basic Skills Program, Final report of the advisory committee. 1978.

Meskauskas, J.A., Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Miller, B.S., (ed.), Minimum Competency Testing. NIE contract No. 400-77-0089 with CEMREL, Inc., St. Louis, Missouri, 1978.

Millman, J., Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Moore, K.K. & Smith, W.B., A rank order approach to discriminant analysis. Proceedings of the Business and Economic Statistics Section of the American Statistical Association, 1975, 451-455.

Nedelsky, L., Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14 , 3-19.

New Jersey Educational Assessment Program Minimum Basic Skills Tests. District Test Coordinator Manual, 1978.

Shepard, L.A., Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 23-32.

Welch, B.L., Notes on discriminant functions. Biometrika, 1939, 31, 218-220.

Zieky, M.L., Communications to the N.J. Minimum Basic Skills committees, 1977.

Zieky, M.L. & Livingston, S.A., Manual for setting standards on the Basic Skills Assessment Tests. Princeton, N.J.: Educational Testing Service, 1977.